How to Manage Your Data

Lisa Spiro

June 2019

This workshop draws heavily on materials from the University of Minnesota Libraries, New England Collaborative Data Management Curriculum, MIT Libraries & DataOne.

Quick Poll: Raise Your Hand If You Have Ever...

- Forgotten what you called a file and/or where you put it
- Discovered unnecessary duplicates, then struggled over which to keep
- Lost data due to hardware failure, lost devices, etc.

Objectives for This Session

1. Understand the importance of managing data.
2. Learn how to create a good data management plan.
3. Name and organize your files effectively.
4. Manage versions of files.
5. Create tidy data.
6. Document your data.
7. Know options for storing, backing up and archiving your data.

1. Why Managing Your Data Matters

What Is Research Data?

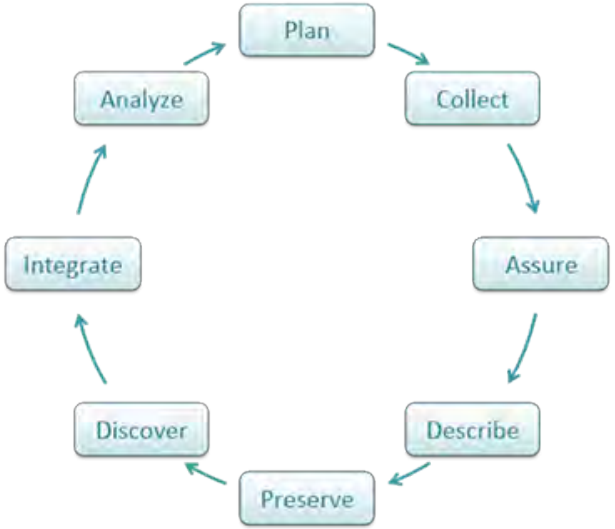
“recorded factual materials that are commonly accepted in the scientific community as necessary to validate research finding” (NSF)

Examples:

- Text, e.g. interview transcriptions & field notes
- Audio & video recordings
- Images
- Numbers, e.g. measurements

What is data management?

The process of storing, organizing, describing, preserving, and sharing data so that research results can be validated, data can be understood, and future use is facilitated.



https://www.dataone.org/best-practices

Why Is Managing Your Data Important?

- Keep track of your data, working more efficiently.
- Prevent data loss.
- Uphold standards of research integrity.
- Make it easier to share and re-use data.
- Meet funder, university & increasingly journal requirements.
- Be kind to Future You and your collaborators.

If the data you need still exists;
If you found the data you need;
If you understand the data you found;
If you trust the data you understand;
If you can use the data you trust;
Someone did a good job of data management.

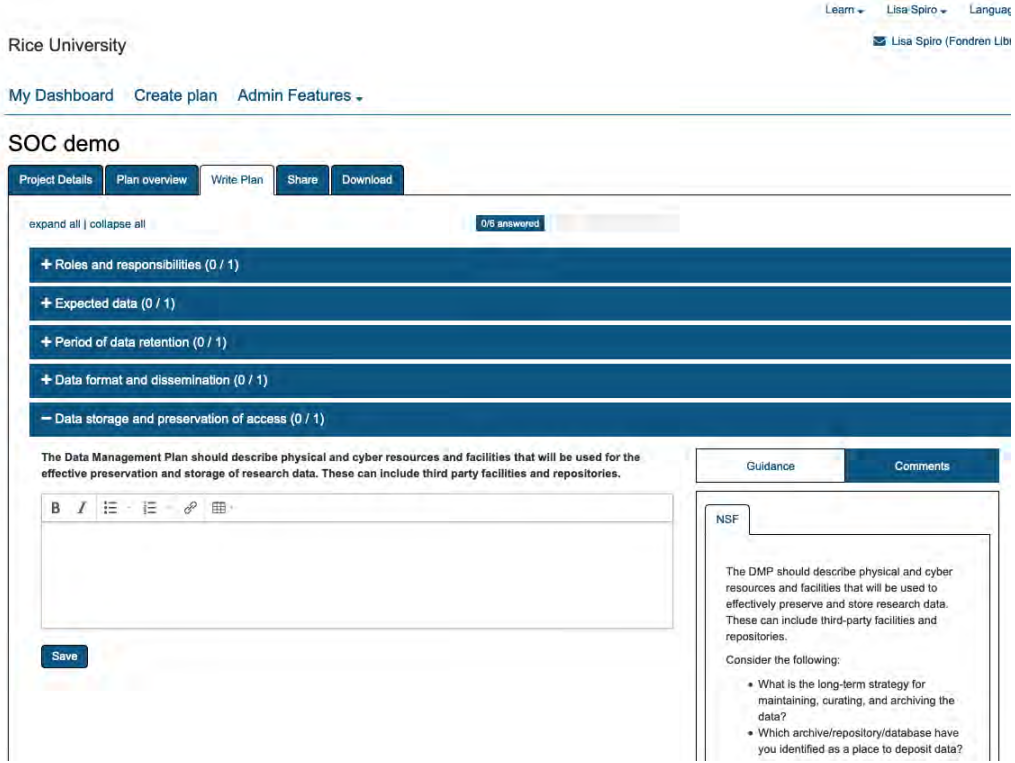
Rex Sanders, USGS

2. Plan

Typical Components of Data Management Plan (NSF)

1. the **types of data** and other materials to be produced in the course of the project;
2. the **standards** to be used for data and metadata format and content;
3. policies for **access & sharing** including provisions for appropriate protection of privacy, security, IP, etc.;
4. policies and provisions for **re-use, re-distribution**, and the production of derivatives; and
5. plans for **archiving** data, samples, and other research products, and for **preservation** of access to them.

Create a Data Management Plan Using DMP Tool



https://dmptool.org/

Key Principles for Data Management Planning

1. Investing time in organizing your data now will save you time later.
2. Be clear and consistent.
3. Work out your data management procedures with collaborators.
4. Document your procedures.
5. Understand that there is no one right way; it's what works for you and your collaborators.

3. Organize Your Data

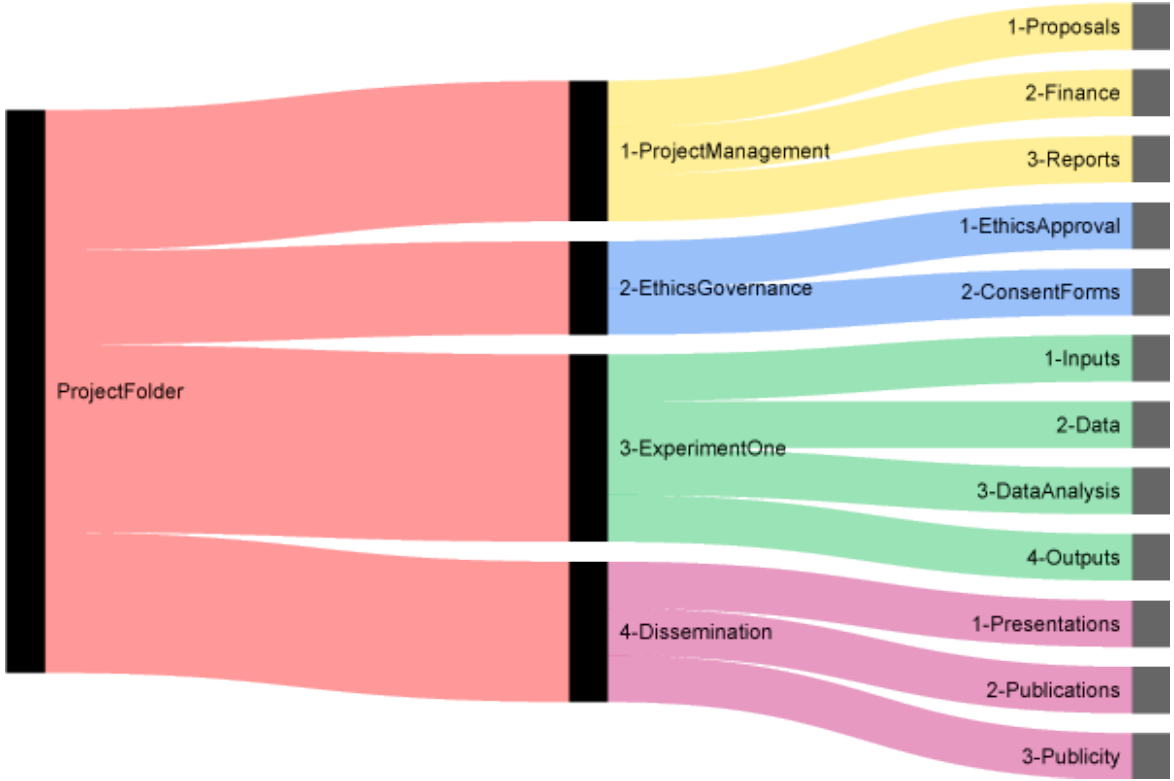
How to Create a Hierarchical File System

1. Organize your files in a predictable, easy-to-sort way.

2. Use relevant categories to organize folders (e.g. Project/Site/Date).

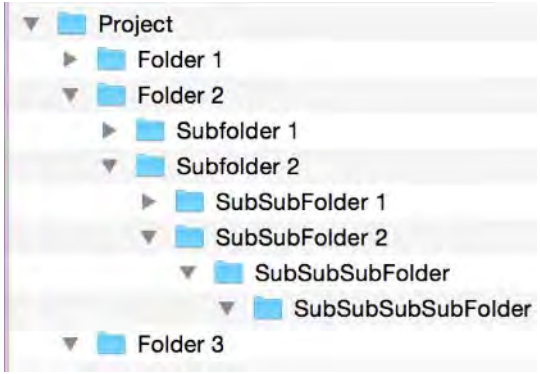
3. Select a meaningful naming convention for folders.

Example of a Directory Structure

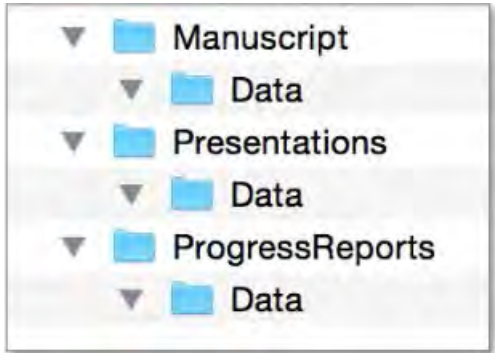


Nikola Vukovic

What to Avoid…



Too much depth



Overlapping categories

Another Option: Tags

Pros:

- More flexible

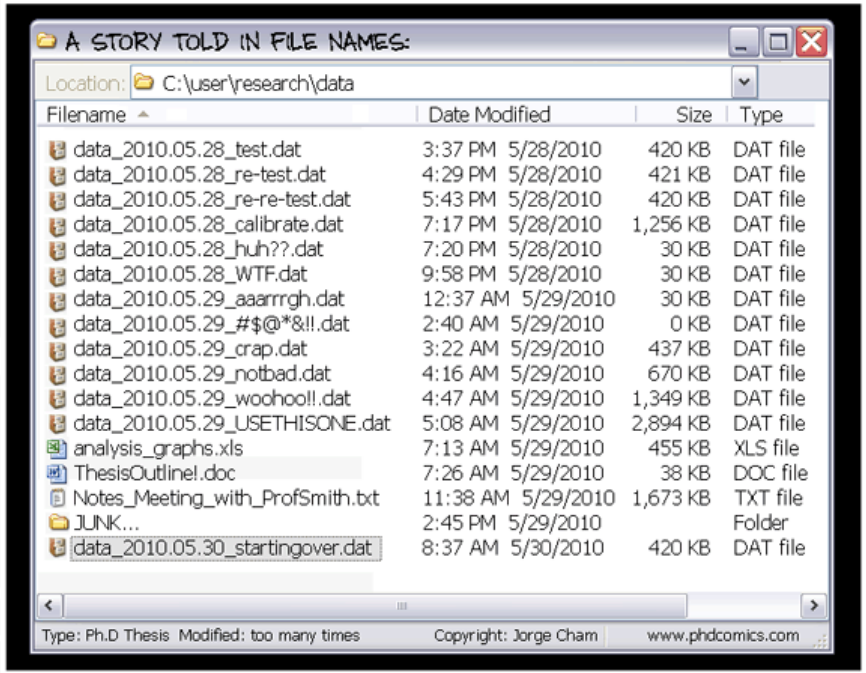
Cons:

- Harder to be consistent.
- May lose tags in moving to different computers.

Systems that support tags:

-Evernote
-Gmail
-Zotero
-Windows 10
-Mac OS
-Box

The Problem of File Names



http://phdcomics.com/comics.php?f=1323

Principles for Effective File Naming

- Files are **distinguishable** from each other within their containing folder.
- Files are easy to **locate, browse** and **sort.**
- If files are moved to another storage platform, their names will retain **useful context**.

(EDINA and Data Library, n.d.) | RDMRose

File Naming Best Practices

- **Be descriptive**: Use shared, meaningful terminology. Incorporate relevant terms such as project name, place, date, experiment, instrument, subject, etc.
  Example: AirQual_Lufkin_Sensor1_201709007

- **Be consistent:** Use the same structure and terms across projects so that files fall into a useful *order* (for sorting) and you can easily identify them.
  Example: AvSAT_Ric_2017
  AvSAT_Ric_2016
  AvSAT_UTx_2017

Guidelines for File Naming

Guideline	Example
Avoid special characters, like / , . # ?	Exp01a.xls, NOT Exp#1.a.xls
Don't use blank spaces. Use CamelCharacters or _ to link together keywords.	Site01_Sensor002, NOT Site1 Sensor 2
Use yyyymmdd for **dates**	200180617, NOT 0617218
Use **leading zeroes**, e.g. 0001, 001, etc	Experiment002.xls, NOT Experiment2.xls

Which file naming scheme works the best?

A. bridgedata1
bridgedata2
bridgedata3

B. bridge1_sensor2_02142013
bridge1_sensor2_02152013
bridge1_sensor2_02162013

C. madisonavebridge_sensor2_20130214
madisonavebridge_sensor2_20130215
madisonavebridge_sensor2_20130216

D. madisonavebridge_sensor2_feb142013
madisonavebridge_sensor2_02152013
madbridge_s2_feb162013

Exercise

Instructions: Review the handout, then partner with 2-3 people to decide on a file naming system in order to **sort by interviewee name**.

3 minutes to discuss

4. Manage versions

Versioning: Which one is authoritative?

DataAnalysis.xls
DataAnalysis2.xls
DataAnalysisSept2017.xls
DataAnalysisFinal.xls
DataAnalysisFinalFINAL.xls

Manual Options for Managing Versions

- Retain original, raw files and significant iterations.
- Use careful file naming: record major changes via whole numbers (v01), minor via an additional number (v02_01)
- Put older versions in an archive folder.
- Create a version control table:

Version Number	Author	Purpose/Change	Date
0-1	Jackie Wilson, Project Manager	Initial draft – to line manager	12/07/2011
0-2	Jackie Wilson, Project Manager	Consultation draft – to working group	21/08/2011
0-3	Jackie Wilson, Project Manager	Second consultation draft – to working group	08/10/2011
1-0	Jackie Wilson, Project Manager	Final version – approved by Project Board	18/11/2011

Software for Managing Versions

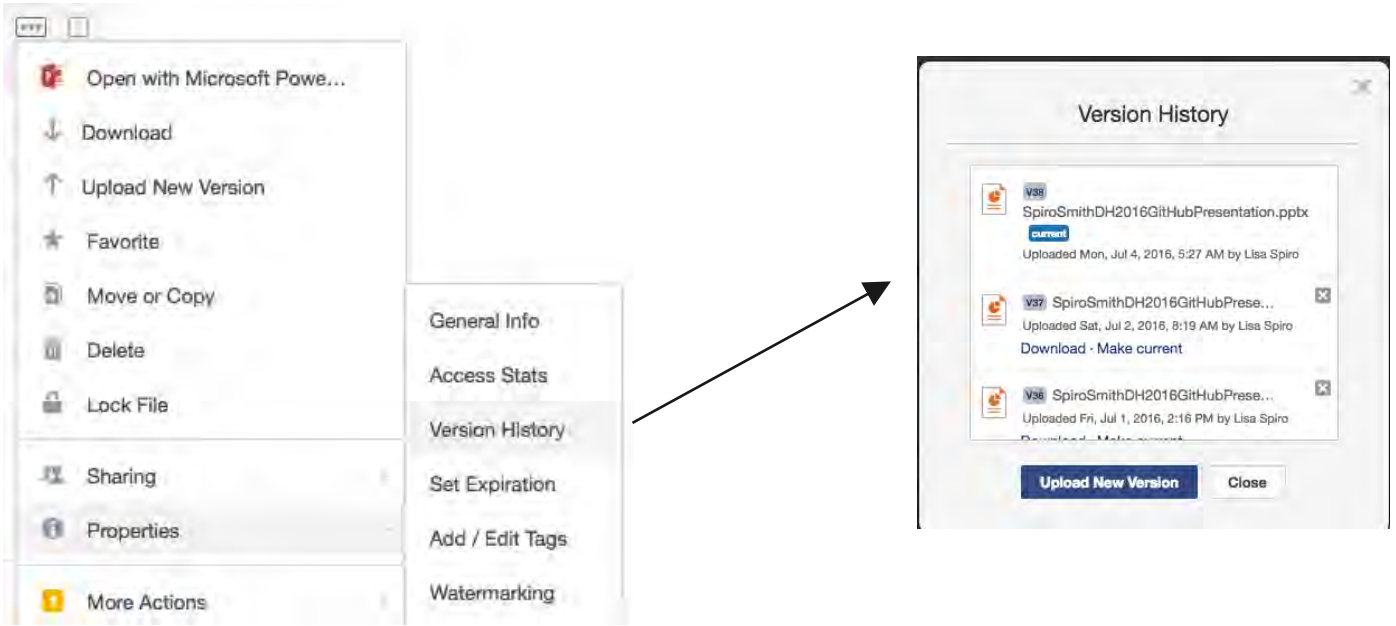
Accessing multiple versions:

- Box, Google Drive & other storage services

Version control software:

- GitHub: Researchers and educators can receive GitHub Team (unlimited repositories) for free.

Accessing Version History on Box.com



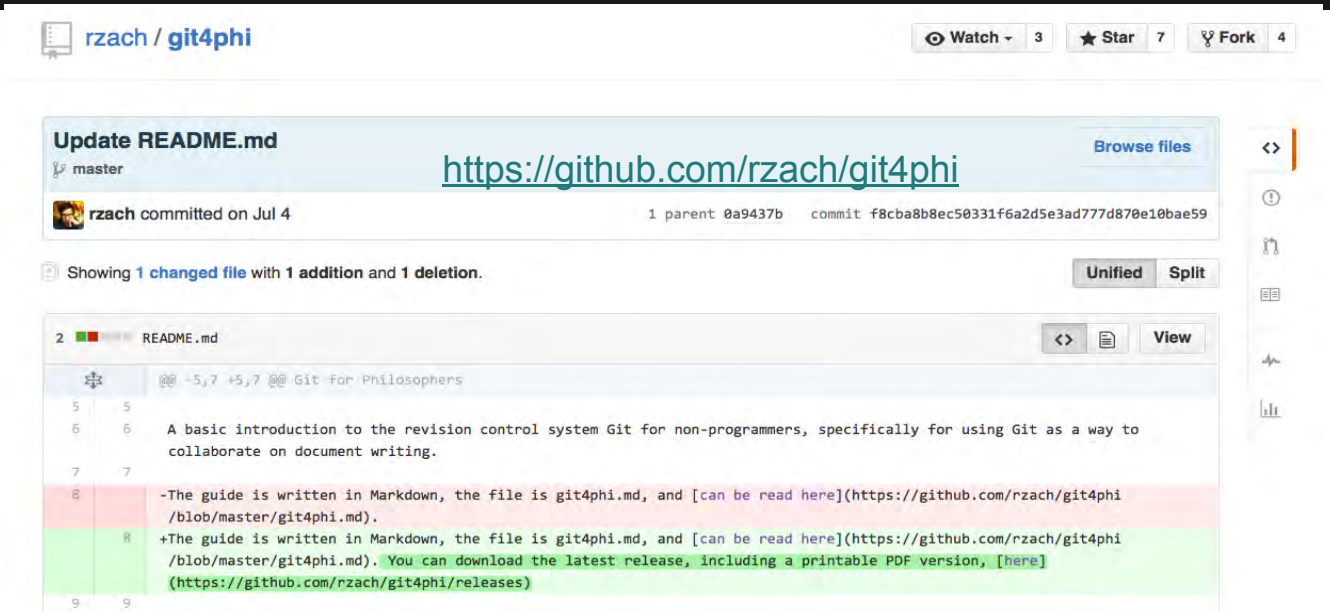
https://community.box.com/t5/Organizing-and-Tracking-Content/Accessing-Version-History/ta-p/50452

Version Control

“Version control is a system that records changes to a file or set of files over time so that you can recall specific versions later.” (Pro Git)

- See who does what.
- Access any version of file.
- Roll back changes.
- Enable new branches of project.

Manage and Access Versions of Files with Git(Hub)



Researchers and educators can receive GitHub Team (unlimited repositories) for free.

5. Create tidy data.

The Problems with Messy Data

- Difficult to analyze
- Requires time to clean
- Confusing to other users– and to Future You
- Raises questions about your credibility

Keep Your Data Tidy

- Make each variable a column & each observation a row
- Make column headers variable names
- Atomize your data; put only a single piece of information in each cell (e.g. city, state, country)
- Be consistent in how you will handle empty values (e.g. NULL, NA, leave blank)

Messy vs. Tidy Data

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

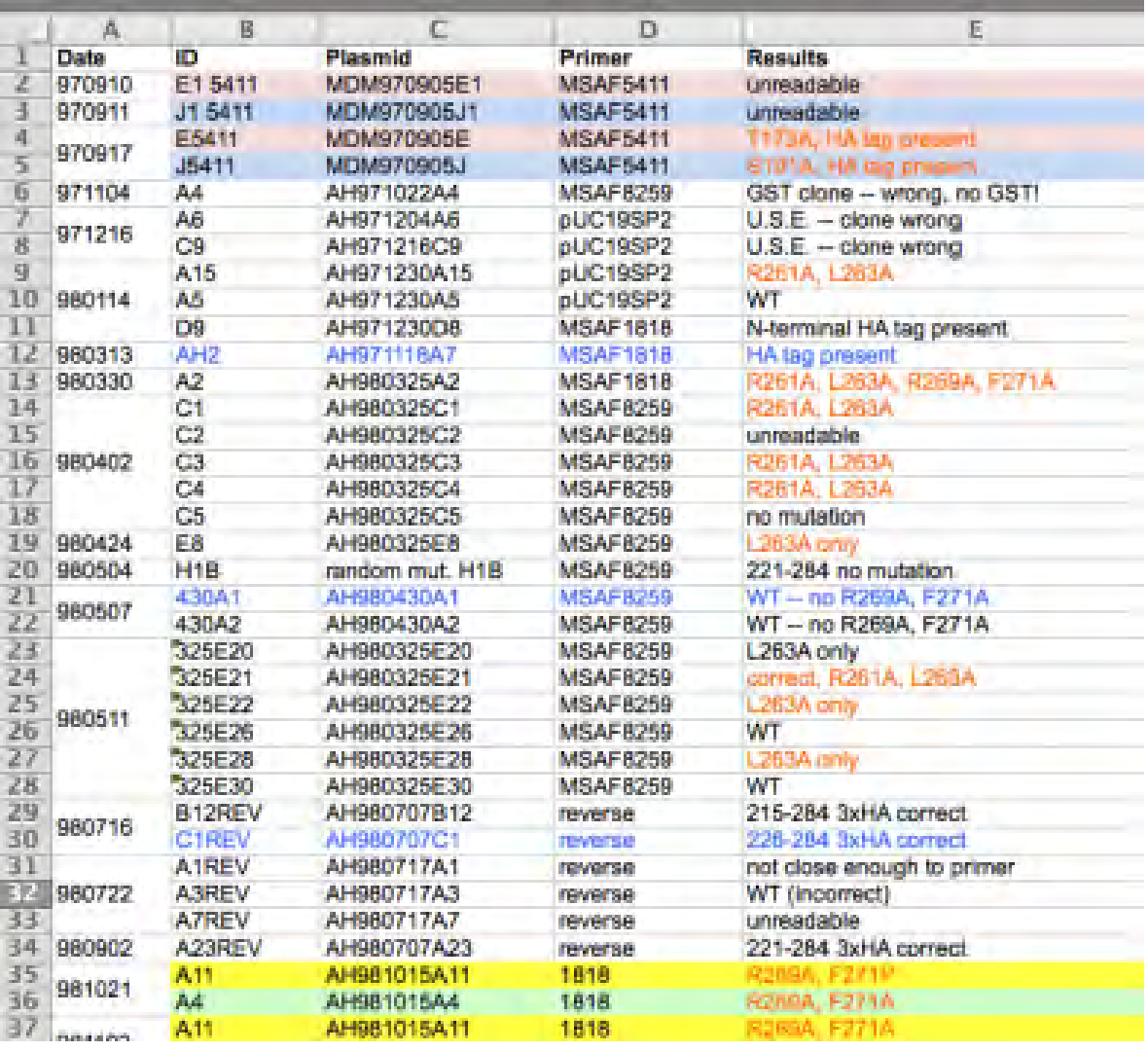
(a) Molten data

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2
AE	2000	m	15–24	4
AE	2000	m	25–34	4
AE	2000	m	35–44	6
AE	2000	m	45–54	5
AE	2000	m	55–64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

(b) Tidy data

Table 10: Tidying the TB dataset requires first melting, and then splitting the `column` column into two variables: `sex` and `age`.

Wickham

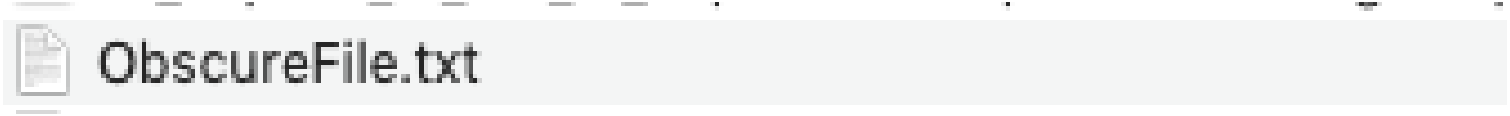


	A	B	C	D	E
1	Date	ID	Plasmid	Primer	Results
2	970910	E1 5411	MDM970905E1	MSAF5411	unreadable
3	970911	J1 5411	MDM970905J1	MSAF5411	unreadable
4	970917	E5411	MDM970905E	MSAF5411	T173A, HA tag present
5		J5411	MDM970905J	MSAF5411	S191A, HA tag present
6	971104	A4	AH971022A4	MSAF8259	GST clone -- wrong, no GST!
7	971216	A6	AH971204A6	pUC19SP2	U.S.E. -- clone wrong
8		C9	AH971216C9	pUC19SP2	U.S.E. -- clone wrong
9		A15	AH971230A15	pUC19SP2	R261A, L263A
10	980114	A5	AH971230A5	pUC19SP2	WT
11		D9	AH971230D8	MSAF1818	N-terminal HA tag present
12	980313	AH2	AH971116A7	MSAF1818	HA tag present
13	980330	A2	AH980325A2	MSAF1818	R261A, L263A, R269A, F271A
14		C1	AH980325C1	MSAF8259	R261A, L263A
15		C2	AH980325C2	MSAF8259	unreadable
16	980402	C3	AH980325C3	MSAF8259	R261A, L263A
17		C4	AH980325C4	MSAF8259	R261A, L263A
18		C5	AH980325C5	MSAF8259	no mutation
19	980424	E8	AH980325E8	MSAF8259	L263A only
20	980504	H1B	random mut. H1B	MSAF8259	221-284 no mutation
21	980507	430A1	AH980430A1	MSAF8259	WT -- no R269A, F271A
22		430A2	AH980430A2	MSAF8259	WT -- no R269A, F271A
23	980511	325E20	AH980325E20	MSAF8259	L263A only
24		325E21	AH980325E21	MSAF8259	correct, R261A, L263A
25		325E22	AH980325E22	MSAF8259	L263A only
26		325E26	AH980325E26	MSAF8259	WT
27		325E28	AH980325E28	MSAF8259	L263A only
28		325E30	AH980325E30	MSAF8259	WT
29	980716	B12REV	AH980707B12	reverse	215-284 3xHA correct
30		C1REV	AH980707C1	reverse	228-284 3xHA correct
31		A1REV	AH980717A1	reverse	not close enough to primer
32	980722	A3REV	AH980717A3	reverse	WT (incorrect)
33		A7REV	AH980717A7	reverse	unreadable
34	980902	A23REV	AH980707A23	reverse	221-284 3xHA correct
35	981021	A11	AH981015A11	1818	R269A, F271A
36		A4	AH981015A4	1818	R269A, F271A
37		A11	AH981015A11	1818	R269A, F271A

What issues do you see with this spreadsheet?

6. Document your data.

What information would you want to know about this file?



Why Document Data?

- Makes it easier for you to interpret your own data
- Facilitates collaboration, sharing, and reuse
- Promotes successful long-term preservation of data

Create a Readme File to Document a File or Directory

Typical contents:

- **What:** title & description
- **When:** date of data collection
- **Who:** name & contact info of creator
- **Where:** location where data was captured
- **How:**
  - Method of data collection, creation or processing
  - Restrictions on accessing files

https://data.research.cornell.edu/content/readme

Files to replicate Sean Bolks and Richard J. Stoll, "The Arms Acquisition Process: The Effect of Internal and External Constraints on Arms Race Dynamics," *The Journal of Conflict Resolution* 44, no. 5 (October 1, 2000): 580–603.

File	Content
table1.dta	Stata data file with data for Table 1
table1.do	Stata .do file with commands to replicate Table 1
table2.dta	Stata data file with data for Table 2
table2.do	Stata .do file with commands to replicate Table

Simple Example of a ReadMe File

Create a Codebook to Describe the Contents of Data Files

“A codebook is an essential document that informs the data user about the **study, data file(s), variables, categories**, etc., that make up a complete dataset. The codebook may include a dataset’s record layout, list of **variable names and labels**, concepts, categories, cases, missing value codes, frequency counts, notes, universe statements, and so on.”
http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook

Codebook Example

CIRP COOPERATIVE INSTITUTIONAL RESEARCH PROGRAM at the HIGHER EDUCATION RESEARCH INSTITUTE AT UCLA

2017 CIRP Freshman Survey (Codebook)

#	Variable Name	Variable Descripion
	ACE	College I.D.
	SUBJID	Subject I.D.
	STUID	Student I.D. as entered on form
	GRPA	Group Code A
	GRPB	Group Code B
1	SEX	Your sex: 1 = Male 2 = Female
2	TRANSGENDER	Do you identify as transgender? 1=No 2=Yes
3	YRGRADHS	In what year did you graduate from high school? 1=2017 2=2016 3=2015 4=2014 or earlier 5=Did not graduate but passed G.E.D. test 6=Never completed high school

Exercise

Think through creating a readme file for one of your datasets (real or imagined) or the “Dr. Psi” data using this template from Cornell.

See “Guidelines for writing ‘readme’ style metadata” http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf

7. Store, Backup and Archive Data

Data Storage Definition

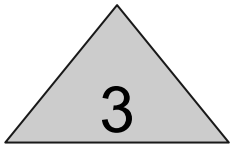
- The media (optical or magnetic) to which you save your data files and software.
- All storage media are vulnerable to risk and obsolescence.
- Storage media should be evaluated and updated every 2-5 years.

Data Backup Definition

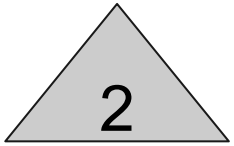
- Allows you to *restore* your data if original data is lost or damaged due to:
  - Hardware or software malfunction
  - Environmental disaster (fire, flood)
  - Theft
  - Unauthorized access

>> TEST YOUR BACKUP PLAN!

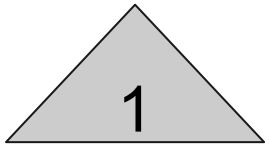
3-2-1 Backup Rule



Save 3 copies of your data.



Use 2 types of storage.



Keep 1 remote copy.

Overview of Data Storage, Backup and Sharing Options at Rice

Network or Cloud Storage

- **storage.rice.edu -** U: drive, departmental shares
- **Research Data Facility (RDF) -** larger scale storage for research

Rice Box: unlimited cloud storage

Backup Options

- **storage.rice.edu** backups/snapshots
- **Crash Plan** for Rice workstations

Data Sharing- Globus Connect

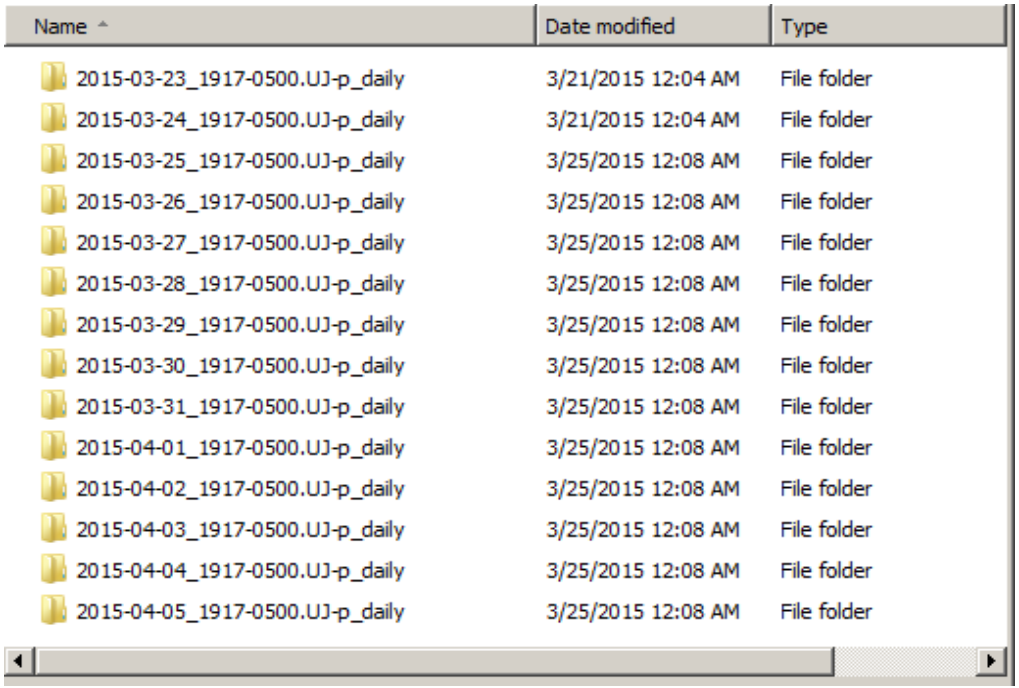
Options for faculty/ staff: https://kb.rice.edu/page.php?id=70762

Options for students: https://kb.rice.edu/page.php?id=65636

storage.rice.edu

- Location: Networked
- Storage quotas
  - Undergraduates: 2 GB
  - Graduates, Staff, Faculty: 5 GB
  - Colleges, Depts, Centers, Institutes: 40 GB
- Performance - Subject to network
- Accessibility
  - NetID folder: Private, not shared
  - Groups: Any Rice NetID holder by request

Accessing your backups on storage



Name	Date modified	Type
2015-03-23_1917-0500.UJ-p_daily	3/21/2015 12:04 AM	File folder
2015-03-24_1917-0500.UJ-p_daily	3/21/2015 12:04 AM	File folder
2015-03-25_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-26_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-27_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-28_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-29_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-30_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-31_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-01_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-02_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-03_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-04_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-05_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder

\\storage.rice.edu\?-home\~snapshot

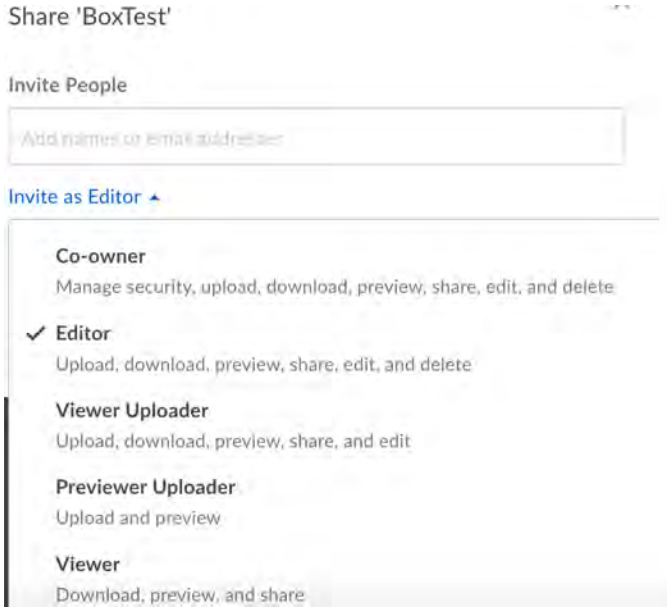
Research Data Facility

- Location: On Site network data shares
- Storage quotas
  - 500GB per researcher
  - Additional storage available with cost recovery
- Performance - Subject to network
- Accessibility
  - Based on NetID and ADRICE security groups
  - Can be shared to multiple users in a research group

Storage Features of Rice Box

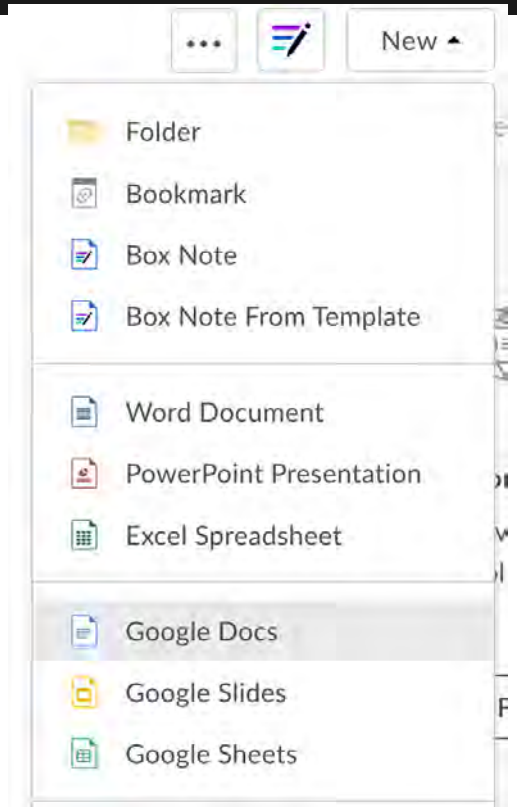
"enterprise cloud-based storage and collaboration service"

- Use unlimited storage (max file size = 15 GB)
- Access prior versions (up to 100)
- Sync files and download for offline use
- Files automatically backed up at multiple data centers
- Control file/folder permissions



Collaboration Features of Box

- Share files, links, notes, etc. with collaborators at Rice & beyond
- Integrate Google Suite
- Take Box Notes and share (can't currently export)
- Use tags to sort and search files (can't currently export)



Security and Rice Box

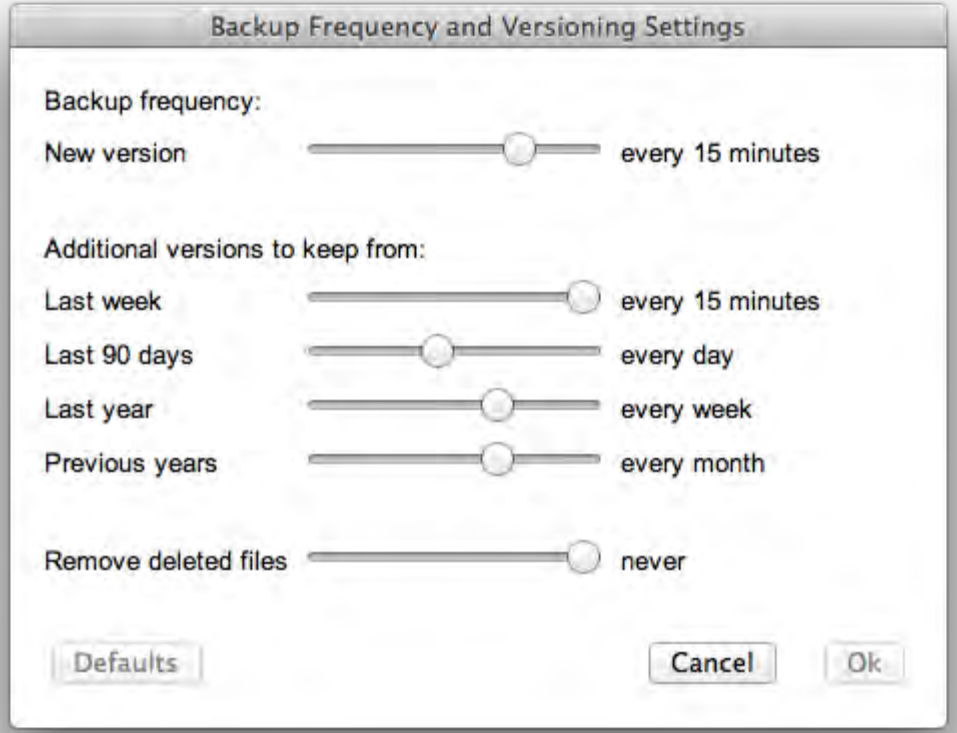
“Box encrypts all data it stores, allow for granular access controls, and facilitates access monitoring. Folders must be properly configured to take full advantage of these security enhancements.”

See Box User Guidelines - Rice Data
Includes recommended settings for sharing, tagging, notifications, and more.
Contact Help Desk about high risk & confidential data.

Backup: CrashPlan

- Availability: Rice-owned computers
- Cost: $82.56/year/person (up to 4 devices)
- Location: Off-site cloud storage
- Procedure: Incremental
- Frequency: Adjustable up to every minute
- Retention: Adjustable up to forever

https://kb.rice.edu/page.php?id=72955



CrashPlan PROe or crashplan.rice.edu

Sharing: Globus Connect

- Widely used service for large data exchange between participating institutions
- Can be used in our HPC environment or from your desktop with Globus Connect Personal
- Accessibility
  - Contact Center for Research Computing to be added to license
  - Arrange for access to peer institution end points

https://www.globus.org/globus-connect

Confidential and Regulated Data (High Risk)

- Health Information, including Protected Health Information
- Health Insurance policy ID numbers
- Social Security Numbers
- Credit card numbers
- Financial account numbers
- Export controlled information under U.S. laws
- Driver's license, passport and visa numbers
- Donor contact information and non-public gift information
- Rice data classified as confidential under policy 808, e.g. birth dates

Securing Your Data

Provide appropriate security for data (e.g. anti-virus protection, access control, encryption, de-identification of data).

Consult IT regarding data security

Approved Services

This table indicates which classifications of data are allowed on a selection of commonly used Rice IT Services.

RICE SERVICE	GENERAL DATA (LOW RISK) POLICY 832	SENSITIVE DATA (MODERATE RISK) POLICY 832 POLICY 808	CONFIDENTIAL DATA (HIGH RISK) POLICY 832 POLICY 808	REGULATED DATA (HIGH RISK) (CUI, HIPAA, PCI) POLICY 832 POLICY 808
Audio and Video Conferencing (Zoom, Camtasia)	✓			
High Performance Computing Research Systems (Spice,HPC Home,Scratch)	✓			
Storage	✓	✓	✓	

https://vpit.rice.edu/it-security/resources/risk-classifications/approved-services

Data Archiving Definition

- Provides a final version of your data.
- Stored for the long-term.
- May be shared publicly through a data archive.

Why Archive Your Research Data with a Data Repository?

- Conform to publisher or funder requirements
- Get cited
  - “studies that made [gene expression microarray] data available in a public repository received 9% … more citations than similar studies for which the data was not made available.” (Piowowar & Vision, 2013)
- Promote future research

Data Archiving Options

Public Repositories:

- Discipline based repository (e.g. GenBank or PANGEA)
- General data repository (e.g. FigShare or Dataverse)
- Institutional repository (e.g. Rice Digital Scholarship Archive)

Private Approaches:

- Long-term storage

Rice Data Sharing Option: Rice Digital Scholarship Archive

https://scholarship.rice.edu/

Data Archiving Caveats

- Do not share confidential data (unless it has been de-identified and approved through IRB).
- Consult with your collaborators before publishing data.
- It may be possible to embargo data so that it is not available until the related publication is released.

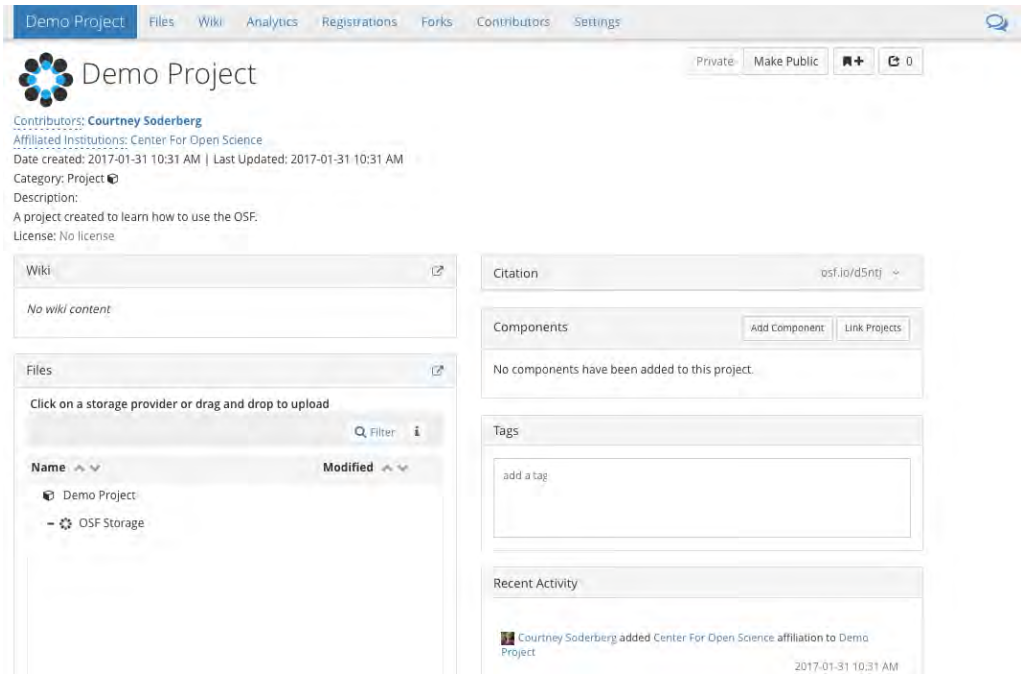
What Does Research Data Services Offer?

https://library.rice.edu/research-data-services

- Workshops on R, Python, Excel, etc. (including upcoming 2 day workshops from Software Carpentry)
- Consulting on finding, analyzing, managing, and visualizing data, including during Friday office hours
- Publishing and preserving data through the Rice Digital Scholarship Archive; providing DOIs
- Reviewing data management plans

Tools and Resources

Use OSF to Manage Your Research Workflow and Collaborate



https://osf.io/

- Organize files in one place
- Share with collaborators
- Control files access
- Integrate with tools like Box
- Track versions
- Make work citable
- Facilitate reproducibility
- Free & open source

Resources

Borer, Elizabeth T., et al "Some Simple Guidelines for Effective Data Management." *Bulletin of the Ecological Society of America* (2009): 205–14.

DataOne Primer on Data Management, https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

Dataverse, *Data Management Plans*, http://best-practices.dataverse.org/data-management/

ICPSR *Guide to Social Science Data Preparation and Archiving,* http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/

Svend Juul et al, "Take good care of your data," http://www.epidata.dk/downloads/takecare.pdf

UK Data Archive, *Managing and Sharing Data: Best Practices for Researchers*, http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Thanks!

Please contact researchdata@rice.edu with any questions.
Visit us online at http://researchdata.rice.edu/.
Help us shape future workshops! Please complete this evaluation: http://library.rice.edu/requests/course-evaluation-form